INSTITUTE FOR DEFENSE ANALYSES

Measures of Effectiveness for Non-Lethal Weapons: Aligning Behavioral Experiments with Operational Success

Shelley Cazares
James Belanich
Jeffrey A. Snyder
P. M. Picucci
Jenny R. Holzer

January 2015
Approved for public release;
distribution is unlimited.
IDA Document NS D-5409
Log: H 15-000028

INSTITUTE FOR DEFENSE ANALYSES
4850 Mark Center Drive
Alexandria, Virginia 22311-1882

About This Publication

This work was conducted by the Institute for Defense Analyses (IDA) under contract HQ0034-14-D-0001, Project DU-2-3825, "Modeling and Simulation of Non-Lethal Weapons (NLW M&S)," for the Joint Non-Lethal Weapons Directorate (JNLWD). The views, opinions, and findings should not be construed as representing the official position of either the Department of Defense or the sponsoring organization.

Measures of Effectiveness for Non-Lethal Weapons:

Aligning Behavioral Experiments with Operational Success

Shelley Cazares, James Belanich, Jeffrey A. Snyder, P. M. Picucci, Jenny R. Holzer

Institute for Defense Analyses, Alexandria, VA

Corresponding Author

Shelley Cazares

Institute for Defense Analyses

4850 Mark Center Drive

Alexandria, VA 22311

Tel: 703 845 6792

Fax: 703 578 2877

Email: scazares@ida.org

Abstract

Non-lethal weapons (NLWs) are weapons, devices, and munitions that are explicitly designed and primarily employed to incapacitate targeted personnel or materiel immediately, while minimizing fatalities, permanent injuries to personnel, and undesired damage to property in the target area or environment. To assess the technical maturity of a NLW system, combat developers compare the system's capabilities to requirements. NLW system requirements could stipulate what physiological degradation the NLW must elicit in the targeted personnel (e.g., temporary visual/hearing impairment in the case of flashbang grenades). Testing such requirements can be straightforward in the laboratory. However, physiology-based requirements can be misleading, since they do not always assess how effectively the NLW can influence the actions of the targeted personnel. Instead, NLW system requirements should, in some cases, stipulate what behavior the targeted personnel must exhibit in response to the NLW. Setting and testing behavior-based requirements is difficult, however, since many factors can influence the targeted personnel's behavior. We developed a framework to guide combat developers in setting and testing behavior-based requirements for NLWs. The framework consists of six main questions to provide structure and discipline for combat developers when determining what behavioral experiments and field data analyses are needed to assess the task effectiveness of specific NLWs in specific military missions. We exercised the framework for a Noncombatant Evacuation Operation scenario, in which U.S. forces consider deploying flashbang grenades against a crowd possibly mixed with para-military forces demonstrating hostile intent. Combat developers could repeat this analysis for other NLWs in other military missions.

Keywords

Non-Lethal Weapon, Less-Than-Lethal Weapon, Flashbang, Crowd Control, Measure of Effectiveness

Measures of Effectiveness for Non-Lethal Weapons:

Aligning Behavioral Experiments with Operational Success

The Department of Defense (DOD) defines non-lethal weapons (NLWs) as weapons, devices, and munitions that are explicitly designed and primarily employed to incapacitate targeted personnel or materiel immediately, while minimizing fatalities, permanent injury to personnel, and undesired damage to property in the target area or environment (DOD, 2013; DOD, 2015a). NLWs are intended to have reversible effects on personnel and materiel (DOD, 2013). Counter-personnel NLWs can potentially deny access to and move, disable, and suppress the targeted personnel (DOD, 2013). Examples of counter-personnel NLWs are sting-ball grenades, human electromuscular incapacitation (HEMI) devices (commonly known as TASERs), dazzling lasers, and flashbang grenades (Joint Non-Lethal Weapons Program (JNLWD), 2013).

A NLW can cause a physiological response in the targeted personnel, such as temporary visual/hearing impairment in the case of a flashbang grenade or temporary inability to move in the case of a HEMI device. Eliciting the physiological response is not the ultimate purpose for deploying the NLW but is rather a means for achieving an operational task, such as preventing the breach of a perimeter fence. The task effectiveness of the NLW must then be considered in the context of the operational mission, such as guarding a building complex. DOD Instruction 5000.02 states that as part of the defense acquisition process, combat developers must assess a NLW system against requirements to estimate the system's technical maturity (DOD, 2015b). Such requirements should quantify the task effectiveness of the NLW, so that the value of the NLW to the warfighter can be extrapolated to other missions with similar tasks.

At first glance, NLW system requirements could be based on the physiological response of the targeted personnel. In the case of a flashbang, these requirements could stipulate the size of the area of temporarily degraded vision elicited in the targeted personnel's field of view and the level of the temporary threshold shift elicited in the targeted personnel's hearing. These requirements can be tested in the laboratory. However, without an understanding of the operational goal of a specific military mission, it can be difficult to set the target thresholds for these physiology-based requirements. For example, what area of degraded vision or what level of hearing threshold shift, if any, would lead to the inability to breach a perimeter fence, if the operational mission is guarding a building complex? Such a question is difficult to answer since measuring physiological responses only begins to assess the effect of the NLW on the targeted personnel. This assessment does not include how the targeted personnel will

subsequently *behave*. For example, a NLW that succeeds at temporarily impairing the vision and/or hearing of the targeted personnel may still not succeed at preventing the targeted personnel from breaching a fence. As such, NLW system requirements based solely on the physiological response cannot assess the task effectiveness of the NLW, since the task must be based on the targeted personnel's behavior. In fact, physiology-based requirements may even be misleading, purporting that a NLW is effective at a task when it has not yet proven to be.

Instead, NLW system requirements should, in some cases, be based on the behavior of the targeted personnel after deployment of the NLW. In contrast with physiology-based requirements, behavior-based requirements can indeed be used to assess task effectiveness. For example, these requirements could stipulate that a flashbang must suppress the targeted personnel from breaching a fence, verbally communicating with an insurgent leader, and/or firing at U.S. forces. Unfortunately, it can be difficult to assess a NLW against behavior-based requirements in a manner that is objective, quantitative, and reproducible since many factors, such as the targeted personnel's training, motivation, and group interactions, influence whether the deployment of the NLW results in the desired change to the targeted personnel's behavior. Furthermore, it can be difficult to define what the desired behavioral change should even *be*.

Many groups have proposed methods for assessing the effectiveness of NLW systems. While these methods define valuable concepts relating to the physiological and behavioral responses of the targeted personnel to a NLW (Kenny et al., 2007; Task Group SAS-060, 2008; Ashworth et al., 2011; Mezzacappa, 2014), they do not provide guidance on how to both *set*, as well as *test*, system requirements for *specific* NLWs used for *specific* military missions. Furthermore, Enclosure 7 of DOD Instruction 5000.02 describes the policies and procedures for human system integration in defense acquisition programs (DOD, 2015b). However, the human involvement is often assumed to consist of only the users, trainers, maintainers, and owners of the system. Rarely is the *targeted personnel* also explicitly considered as part of the system. Fortunately, many concepts already called for by DOD Instruction 5000.02 can be repurposed to explicitly consider the targeted personnel.

Building upon the concepts already set forth by others (Kenny et al., 2007; Task Group SAS-060, 2008; Ashworth et al., 2011; Mezzacappa, 2014; DOD, 2015b) we developed a framework to guide combat developers through the challenges in both setting and testing behavior-based system requirements for specific NLWs in specific military missions. In this paper, we define our framework and exercise it for a particular military scenario: a

Noncombatant Evacuation Operation (NEO). We then discuss how combat developers could use our framework to estimate the task effectiveness of other NLW systems for other military missions.

Method

Our framework consists of six main questions:

1. What is the scenario?
2. What are the constraints of the scenario?
    a. What is the operational goal of the mission?
    b. What are the Rules of Engagement (ROE)?
3. What actions could the targeted personnel take that are relevant to the scenario? Relevant actions must
    a. Potentially thwart the operational goal of the mission and
    b. Be within the window of opportunity for the weapon.
4. What metrics describe how the weapon influences the relevant actions?
5. What behavioral experiments must be done to acquire the desired metrics?
    a. From what pool should the experimental participants be drawn?
    b. What instructions and training should the participants be given before the test?
    c. What steps should the test consist of?
    d. What instrumentation is needed to measure the participants' actions during the test?
    e. How should the collected data be analyzed to calculate the desired metrics?
    f. What safety constraints might be imposed?
6. What field data are available to estimate the desired metrics?

Our framework allows combat developers to determine which experiments and field data analyses will be most relevant for setting and testing behavior-based requirements for NLW systems. Questions #1–4 guide combat developers in *setting* the requirements. When answering these questions, combat developers should solicit input from military operators who have relevant experience with NLWs. Questions #4–6 guide combat developers in *testing* a NLW system against its behavior-based requirements. When answering these questions, combat developers should solicit input from NLW researchers who design and execute human behavioral experiments. Question #4 is tricky since its answer can guide both the setting and testing of behavior-based requirements and, as such, can benefit from input from both NLW operators and researchers. Combat developers may find that the input received

from operators is at odds with the input received from researchers. Our framework provides combat developers the opportunity to identify and reconcile any such discrepancies early in the NLW system's defense acquisition process. Furthermore, the framework fits in well with concepts called for in DOD Instruction 5000.02 (DOD, 2015b).

Discussion

To provide an example of how our framework's questions can be answered, we exercised our framework for a NEO scenario. The DOD defines NEOs as operations that are "directed by the Department of State or other appropriate authority, in conjunction with the DOD, whereby noncombatants are evacuated from foreign countries when their lives are endangered by war, civil unrest, or natural disaster to safe havens" (DOD, 2015a, p. 177). NEOs occurred in Somalia in 1991 (Siegel, 1991) and Lebanon in 2006 (Government Accountability Office, 2007).

Question #1: What is the scenario?

In Question #1, combat developers define a detailed military scenario in which NLWs could be used against targeted personnel. The NEO scenario selected for this exercise is similar to a scenario posed in a U.S. Marine Corps (USMC) Concepts of Operations document (USMC, 2011) and can be summarized as follows:

> An undeveloped country is embroiled in civil conflict. Some factions are openly hostile to the U.S. while others seek support from Western powers. U.S. forces are tasked with the protection of the U.S. embassy and its adjacent helicopter-landing zone. These forces must use various methods, equipment, and weapons including flashbang grenades to keep restricted areas clear of local nationals who are not being evacuated and protect the embassy against hostile actions while minimizing non-combatant casualties and collateral damage. The primary threat consists of para-military factions and criminal elements that seek to exploit the ongoing civil unrest. Some para-military factions will attempt to inflict casualties on U.S. personnel if given a low-risk opportunity. In addition to the armed factions, crowds armed only with rocks and sticks have demonstrated near several embassies and have entered and ransacked other areas that are not well protected. The para-military forces sometimes hide themselves within the crowds, making them difficult to detect. Shortly after the U.S. forces arrive, airborne surveillance systems detect a crowd forming and moving toward the embassy. The crowd is diverse, including some women with children. Initial indicators suggest that a few men armed with AK-47 type rifles may be

intermixed. Evacuation operations are in progress with helicopters cycling through the landing zone and several dozen evacuees waiting for processing.

Question #2: What are the constraints of the scenario?

In Question #2, combat developers analyze the selected scenario to determine its constraints. Two separate concepts must be considered. First, the operational goal of the mission must be clearly identified. The goal of our NEO scenario is the safe evacuation of the noncombatants from the embassy. Second, the ROE must be explicitly stated or otherwise inferred. The DOD defines ROE as "directives issued by competent military authority that delineate the circumstances and limitations under which U.S. forces will initiate and/or continue combat engagement with other forces encountered" (DOD, 2015a, p. 217). In our NEO scenario, consideration of the ROE leads to questions such as: Are the U.S. forces allowed to use non-lethal and lethal force against the crowd? How can force escalate via graduated response? (The details of force escalation are often provided underneath the ROE, such as in a graduated response matrix (DOD, 2007). For the sake of brevity, however, the entirety of these rules are subsumed in our use of the term *ROE*.) In our NEO scenario, the ROE have been loosely defined, stating only that non-combatant casualties and collateral damage should be minimized.

Question #3: What actions could the targeted personnel take that are relevant to the scenario?

In Question #3, combat developers determine the relevant actions that the targeted personnel could take. In our NEO scenario, the demonstrating crowd could perform many different actions while approaching the embassy. Only some of these actions are relevant, however. Relevant actions must meet two criteria.

First, a relevant action must potentially thwart the goal of the mission (i.e., the safe evacuation of the noncombatants). Other actions of the crowd, such as banging on the embassy fence and shouting anti-American slogans, may be undesired by the U.S. forces but will not prevent the safe evacuation of the noncombatants. Therefore, there is little reason to assess the capability of a NLW to suitably influence those actions since they are not operationally relevant to the mission.

Second, a relevant action must also fall within the window of opportunity for non-lethal force. In our NEO scenario, the ROE dictate that noncombatant casualties and collateral damage to the crowd must be minimized, but not necessarily disallowed. As such, there may be some extreme actions that the crowd could take for which lethal force is indeed authorized. An example of an extreme action is preparing and aiming a rocket-propelled grenade toward the embassy. Such an action would clearly indicate hostile intent to immediately do harm to the evacuees and

U.S. forces. There is little reason to assess the capability of a NLW to suitably influence such an extreme action since the ROE would likely authorize lethal, rather than non-lethal, force in response.

The first column of Table 1 lists relevant actions for the NEO scenario. “Climb over fence,” “aim and throw rock,” “aim and fire rifle,” and “verbally pass message” are identified as relevant actions in the NEO scenario. These actions encompass the most basic needs of a potential adversary to move, shoot, and communicate.

All four actions could potentially thwart the safe evacuation of noncombatants from the embassy. Of course, rifles aimed and shot at evacuees or U.S. forces could thwart the evacuation. Furthermore, although one rock aimed and thrown at evacuees or U.S. forces or one demonstrator climbing over the embassy fence may not thwart the evacuation, several rocks thrown or several climbers may. Finally, while speech between some demonstrators may not thwart the evacuation, para-military command and control via verbal orders (e.g., “Storm the second gate on the right!”) could lead directly to other actions that may thwart the evacuation.

All four actions also fall within the window of opportunity for non-lethal force. The ROE are unlikely to authorize lethal force against a demonstrator simply for climbing a fence, throwing a rock, or speaking with others. “Aim and fire rifle” could also fall into the non-lethal window of opportunity. The ROE may not authorize lethal force until after the first shot has been fired. NLWs could help delay or even prevent that first shot.

Question #4: What metrics describe how the weapon influences the relevant actions?

In Question #4, combat developers define metrics to assess the capability of the NLW to suitably influence the relevant actions. These metrics are the very same ones on which the NLW system requirements should be based. Combat developers must consider the operational relevance of each proposed metric as well as the constraints of collecting data to estimate each one. Military operators of NLWs in theater can provide input about which proposed metrics are most operationally relevant. Researchers can offer suggestions on which proposed metrics can be obtained within the constraints of experimental logistics, cost, safety, and ethics.

Metrics for behavioral experiments. Several potential metrics could be used to assess the capability of a NLW to suppress the crowd’s fence-climbing action. However, some metrics are better than others for quantifying the NLW’s effectiveness at this task. For example, one could measure the average speed (e.g., in meters per second) at which participants climb a fence in a behavioral experiment. This speed could be measured immediately after the detonation of a flashbang (or a flashbang surrogate used to meet experimental safety constraints, as discussed below). One could also measure climbing speed without a flashbang/surrogate. A reduction in speed with versus

without the flashbang/surrogate could indicate the extent that the flashbang can suppress fence-climbing behavior. However, this metric has little operational relevance. That is, potential consumers of flashbangs (i.e., NEO commanders) will require an immediate and intuitive understanding of what a flashbang can "buy" them in theater. For this reason, alternative metrics should be considered instead. As an example, one could measure the time it takes before *N* participants in a multi-participant behavioral experiment get a foot over the top of the fence, as listed in the second column of Table 1. An increase in time with versus without the flashbang/surrogate could indicate the extent of the flashbang's suppressive effect. This metric is more operationally relevant since it describes how much time a flashbang can "buy" the guards in the NEO scenario.

The second column of Table 1 also lists the metric for the rock-throwing action. Although average miss distance is often used to assess throwing actions, it does not have as much operational relevance at it may initially appear, since it can be difficult to immediately intuit what an increase in miss distance can "buy" the U.S. forces in the NEO scenario. For example, a thrown rock will fail to harm its intended target, regardless of whether it misses by 10 cm or 10 m. Instead, one could measure the number of rocks "accurately" thrown in a particular time window (e.g., 5 min). The definition of "accurate" in this context is based upon the potential for damage caused by a thrown rock and the U.S. forces' expected response based on the ROE. An "accurately" thrown rock is defined here as one that hits the head of an evacuee or member of the U.S. forces, since hitting any other part of the body with a rock is not likely to cause enough damage to thwart the evacuation. This metric could therefore succinctly describe to a potential consumer just how many "accurately" thrown rocks the flashbang can prevent in the NEO scenario.

One could also measure the time until the first "accurate" rifle shot, as per the second column of Table 1. Unlike with a rock, a rifle shot that does not hit its intended target but passes nearby might still lead to an escalation of force, as per the ROE in the NEO scenario, which, in turn, could thwart the safe evacuation of the noncombatants. Still, under the ROE envisioned in our NEO scenario, a rifle that is shot far from evacuees or U.S. forces will likely fail to elicit an escalation of force, regardless of whether it misses an evacuee or U.S. force member by 10 m or 100 m. As such, an "accurate" rifle shot is defined here as one that passes within 1 m of an evacuee or U.S. force member. Furthermore, since the first "accurate" rifle shot could lead to an escalation of force, there is no point in counting the total number of "accurate" rifle shots within a time window. In contrast to rock throws, only the first "accurate" rifle shot is guaranteed to be unbiased by any subsequent escalation in force in the NEO scenario. As

such, our selected metric for the rifle-shooting action describes how much time a flashbang can "buy" the U.S. forces in the NEO scenario before "accurate" rifle fire is first received.

Finally, the verbal message-passing metric could be estimated by verbally asking the participants a series of pre-written, easily answerable questions. Each participant's answer could be compared to an "answer key." A reduction in the percent of questions answered correctly with versus without a flashbang/surrogate could indicate the flashbang's capability to suppress verbal message-passing.

Metrics for field data analyses. The previous discussion has focused on metrics estimated in behavioral experiments. However, some experiments, such as those involving large crowds or potentially harmful NLWs, may not always be feasible due to constraints in logistics, cost, ethics, and safety. Therefore, some metrics may have to be estimated from existing data of NLWs used in theater. The third column of Table 1 lists metrics for field data analyses. Experimental versus field data metrics differ more for some actions than for others. For example, a similar rifle-firing metric estimated in an experiment ("time until first 'accurate' shot") could also be estimated from field data, using the time at which the NLW was deployed as the starting time for measurement. Slightly different metrics should be selected for the rock-throwing action, however. While the experimental metric is "number of 'accurate' throws within a time window," the field data metric is "time until the first 'accurate' throw." Depending on the actual ROE that were in effect when the field data were collected, it may be that force was indeed escalated after only some or even one rock was "accurately" thrown. Therefore, any estimate of the total number of "accurate" throws could be biased by any escalation in force. A similar bias could also affect the field data metric for the fence-climbing action. As a result, "time until the first demonstrator makes it to the top of the obstacle" should be estimated from field data, as opposed to the "time until *N* participants put their first foot over the top of the fence." Finally, the field data metric for verbal message-passing differs widely from its experimental counterpart. An experiment can control for the types of questions the participants are asked, such that the participants' receipt of the questions can be scored based upon their answers. In field data, however, the content of the messages is likely unknown, and the participant's receipt of the messages cannot be scored in a similar way. Other indicators of verbal message-passing could be used instead, particularly if video footage is available. Such indicators include a reduction in the number of conversations between demonstrators or the demonstrators' increased use of alternative communication methods (e.g., hand gestures) immediately after the detonation of a flashbang.

Question #5: What experiments can be done to acquire the desired metrics?

In Question #5, combat developers build a roadmap describing what experiments could and should be conducted to collect data from which the selected metrics can be estimated. A series of more than one experiment may be needed to investigate each metric. The first experiment should have high *internal validity*; it should consist of few to no uncontrolled factors. Changes in a dependent variable (e.g., a metric selected in Question #4) can therefore be attributed directly to changes in the independent variable(s) (e.g., the dose of flash and bang). However, an experiment with high internal validity may not generalize well to the real world, since so many factors are held constant (Shaughnessy, Zechmeister, & Zechmeister, 2011). Therefore, later experiments should strive for higher *external validity*; more factors should be left uncontrolled so that the experimental results generalize better to the real world. Of course, the more uncontrolled factors, the more difficult it will be to attribute the dependent variable (e.g., a Question #4 metric) to any one particular independent variable (Shaughnessy et al., 2011). However, the body of knowledge built up in the first few experiments can help investigators make this leap.

Early experiment: Single participant for high internal validity. The first experiment to collect data for the fence-climbing metric can be straightforward, such as in previous experiments (McLin et al., 2010). In our envisioned experiment, participants could be drawn from a pool of healthy adults who are similar in skill to both the para-military members and other demonstrators in the NEO scenario. Participants must meet inclusion and exclusion criteria for enrollment in the experiment. These criteria may state that each participant must be a healthy adult who, under normal conditions, can climb a fence unaided. Each participant must be given instructions and training before the start of each test. The participants should not be instructed to "think and act like a para-military member" or a "demonstrator," since the participant is unlikely to know how a para-military member or demonstrator thinks or acts; instead, the instructions should motivate the participant to complete the relevant actions (Mezzacappa, 2014). For example, the participant could be instructed that he or she will earn a small reward (e.g., $10) if he or she makes it to the other side of the fence within a particular time window. Before the test, the participant could practice climbing the fence for a particular length of time to ensure that he or she has reached a particular level of capability.

The test and subsequent data analysis should consist of a straightforward series of steps. For example, immediately before the test, a single participant could be placed in front of the fence, and a buzzer could indicate the start of the run. Three-dimensional video could track the location versus time of reflectors attached to the participant's shoes. This test could be repeated several times, resulting in several runs for each participant. Some runs could serve as test runs, where a flashbang/surrogate is deployed soon (e.g., 1.5 sec) after the buzzer. Other

runs could serve as controls, where no flashbangs/surrogates are deployed. For each run, the time could be measured between the buzzer sounding and the participant's shoe surpassing the top of the fence. A statistical analysis could then be performed to determine whether the metric selected in Question #4 is statistically different with versus without the flashbang/surrogate. The experiment must be powered to ensure that any statistically significant difference is also operationally significant. Sample size calculations could be done during the design of the experiment to determine how many participants must be enrolled and how many runs each participant must complete to achieve the necessary level of statistical power (Gardiner, 1997). Other straightforward experiments could collect initial data for the rifle-shooting, rock-throwing, and verbal message-passing metrics, such as in previous experiments (Mullins & Limberg, 2007; DeMarco, Reid, Tevis, Chua, & Riedener, 2010; Mezzacappa, 2014; Rahimi, Borve, & Arnesen, 2013).

NLW surrogates for experimental safety. There is a risk that exposure to some types of NLWs can cause permanent injuries to experimental participants. To reduce that risk, some experiments could make use of surrogates for NLWs. For example, flashbang surrogates could include masking doses of light and sound that are bright and loud enough to interfere with the participant's sensory input but are not bright or loud enough to cause permanent damage to the participant's sensory organs (Thys, 2000). Other types of surrogates could be used to block the sensory input at the same physiological level as a flashbang would. For example, during test runs, the participants could wear electronic visors and headphones that produce the same magnitude and type of temporarily degraded vision and hearing as a flashbang that was deployed at a particular range and orientation to the participant, as determined through previous experimentation (human or animal) or modeling and simulation (M&S). A lower-cost but also lower-fidelity alternative is to use plastic goggles with a particular area of the lenses blacked out and foam earplugs rated to a particular decibel level of hearing degradation (Sauerburger, 1998), each of which can be purchased for less than $10 at a local hardware store. Although the use of NLW surrogates could come at the cost of lower external validity, their use could ease the path to achieving safety approval for the experimental protocol.

Later experiment: Multiple participants for high external validity. Another experiment could involve multiple participants simultaneously, such as in previous experiments (Muir, Marrison, & Evans, 1989; Mullins & Limberg, 2007; Mezzacappa, 2014). In our envisioned experiment, different levels of surrogates could be given to different participants, and these surrogates could mimic the effect of a flashbang deployed closer to some participants than to others. Different levels of motivation (e.g., monetary rewards) and training (e.g., time spent

practicing to climb the fence) could be applied to different participants, who could be binned into low-, moderate-, and high-motivation/training groups. In addition, one participant could be told that he or she will earn a reward each time another participant completes his or her task. This incentive could motivate that participant to instigate those behaviors in others, potentially eliciting behavior similar to that of para-military leaders in the NEO scenario.

Final experiment: Determination of intent. A final experiment could assess the capability of the flashbang to assist U.S. forces in determining the intent of the targeted personnel. Regardless of whether the flashbang/surrogate is able to suppress the targeted personnel's fence-climbing, rock-throwing, rifle-shooting, or verbal message-passing behaviors, the use of the flashbang/surrogate may still elicit enough information to determine the targeted personnel's intent. Such a result could help refine the tactics, techniques, and procedures (TTP) for the use of flashbangs in NEOs (DOD, 2007). A video of the multi-participant experiment described previously could be shown to new participants serving in the role of the U.S. forces guarding the embassy in the NEO scenario. These new participants could be drawn from a pool of active duty military personnel, and inclusion criteria could ensure that each has guarded an embassy within the past year. Before the test, each military participant could be told that he or she will earn a reward if he or she correctly rates the motivation/training bin (i.e., low, medium, high, or instigator) of at least 90% of the individuals in the video. While watching the video, each military participant could write down the identification numbers of which videoed participants are believed to fall into which bin. Afterwards, investigators could perform a statistical analysis to determine whether the military participants' ratings were statistically and operationally different from truth.

Question #6: What field data are available to estimate the desired metrics?

In Question #6, combat developers devise a similar roadmap for field data analyses in those cases for which behavioral experimentation is not possible. To our knowledge, few to no field data exist concerning the use of flashbangs in NEOs. (In fact, multiple discussions within the NLW community have revealed that flashbangs are mostly used by the military in room or building clearance operations.) Video footage of detainee riots in Operation Iraqi Freedom (Orbons, 2012) may be a viable source of field data for flashbangs in crowd control. Flashbangs were deployed against rioting detainees as they approached the prison fence but before they reached it. As such, no detainees attempted to climb the fence, and therefore it is not possible to estimate the fence-climbing metric listed in the third column of Table 1. Furthermore, the detainees did not have access to rifles; therefore, the rifle-shooting metric cannot be estimated either. However, the detainees did throw rocks and other objects at the prison guards, and

therefore it is possible to estimate the rock-throwing metric. Finally, it may be possible to estimate the verbal message-passing metrics listed in the third column of Table 1, including a qualitative assessment of whether, immediately after deployment of the flashbangs, conversation ceased among the detainees and/or the detainees exhibited an increased use of hand gestures.

Conclusion

We created our framework to assist combat developers in assessing the task effectiveness of NLWs by setting and testing appropriate behavior-based requirements for NLW systems. Although NLWs cause a physiological response in the targeted personnel, that response is not the ultimate purpose for deploying the NLW. The ultimate purpose is to allow U.S. forces to succeed at the operational tasks relevant to their mission. Exercising this framework for the NEO scenario provided an example of how the framework could be used to determine which behavioral experiments and field data analyses are needed to estimate the task effectiveness of a flashbang for a particular military crowd control mission. Combat developers could repeat this analysis for other NLW systems in other military scenarios, such as dazzling lasers used at vehicle checkpoints. Combat developers could then incorporate their answers to the framework's questions into Statements of Work (SOWs) for performers who design and execute behavioral experiments and field data analyses. These SOWs could provide direction to performers, helping to ensure that their efforts provide the information needed to set and test NLW system requirements.

Our framework could also be used in other ways. First, although the framework was created for NLWs, it could be used for any weapon, including lethal weapons. Second, the framework focuses on behavioral experiments and field data analyses, both of which can be constrained by time and funding. DOD Instruction 5000.02 supports M&S as an additional approach to test and evaluation (DOD, 2015b). Thus, the framework could be leveraged to guide NLW M&S efforts. Finally, the framework was created with NLW system requirements in mind. System requirements specify how well the system must perform *when used as intended.* Determining how a NLW system should best be used is a complex topic. The framework could be leveraged to investigate the appropriate TTP for NLWs and the methods needed to train military operators to employ those TTP in a military mission.

References

Ashworth, A., Simonds, J., Fleming, J., Van Hove, J., Dayton, T., & Kohn, N. (2011). *Status of the US Air Force non-lethal weapons human effects-based design program*. Paper presented at the 6th European Symposium on Non-Lethal Weapons, Ettlingen, Germany.

DeMarco, R. M., Reid, G., Tevis, K., Chua, F. B., & Riedener, J. (2010). *Automated targeting systems to assess armed human response to non-lethal threats.* Paper presented at the 27th Army Science Conference, Orlando, FL.

Department of Defense. (2007). *NLW multi-service tactics, techniques, and procedures for the tactical employment of nonlethal weapons* (FM 3-22.40, MCWP 3-15.8, NTTP 3-07.3.2, AFTTP(I) 3-2.45). Washington, DC: Army, Marine Corps, Navy, & Air Force. Retrieved from https://www.fas.org/irp/doddir/army/fm3-22-40.pdf

Department of Defense. (2013). *DoD executive agent for non-lethal weapons (NLW), and NLW policy* (Department of Defense Directive 3000.03E). Washington, DC: Undersecretary of Defense for Acquisition, Technology, and Logistics. Retrieved from http://www.dtic.mil/whs/directives/corres/pdf/300003p.pdf.

Department of Defense. (2015a). *Department of Defense dictionary of military and associated terms* (Joint Publication 1-02). Washington, DC: Author. Retrieved from http://www.dtic.mil/doctrine/new_pubs/jp1_02.pdf

Department of Defense. (2015b). *Operation of the defense acquisition system* (Department of Defense Instruction 5000.02 (Interim)). Washington, DC: Undersecretary of Defense for Acquisition, Technology, and Logistics. Retrieved from http://www.acq.osd.mil/docs/DSD%205000.02_Memo+Doc.pdf

Government Accountability Office. (2007). *State Department: The July 2006 evacuation of American citizens from Lebanon* (GAO-07-893R). Washington, DC: Author. Retrieved from http://www.gao.gov/assets/100/94959.pdf

Gardiner, W. P. (1997). *Statistics for the biosciences*. London, United Kingdom: Prentice Hall.

Joint Non-Lethal Weapons Program. (2013). *Non-lethal weapons annual review.* Quantico, VA: Author. Retrieved from http://jnlwp.defense.gov/Portals/50/Documents/Press_Room/Annual_Reviews_Reports/2013/DoD_Non-Lethal_Weapons_Program_Annual_Review_11.19.2012_HTML_format_v1.pdf

Kenny, J. M., Scholl, D., Murray, B., Farrer, D., Sokolowski, J., Dolan, D., Peters, D., McShea, L., & Finch, K. (2007). *Establishing a framework to determine non-lethal weapon effectiveness*. State College, PA: Penn State Applied Research Laboratory.

McLin, L. N., Foutch, B. K., Hasbrouck, J. T., Chauvin, D. J., Kuyk, T. K., Smith, P. A., Stringham, J. M., & Garcia, P. V. (2010). *Warning and suppression with single and multiple laser optical warning devices in a*

field setting. Brooks City-Base, TX: Air Force Research Laboratory, 711 Human Performance Wing, Human Effectiveness Directorate, Directed Energy Bioeffects, Optical Radiation Branch.

Mezzacappa, E. (2014). Effectiveness testing of non-lethal weapons. *The Journal of Defense Modeling and Simulation: Applications, Methodology, Technology*, September 11. doi:10.1177/1548512913502880

Muir, H., Marrison, A., & Evans, C. (1989). *Aircraft evacuations: the effect of passenger motivation and cabin configuration adjacent to the exit* (CAA-P 89019). London, United Kingdom: Civil Aviation Authority.

Mullins, M., & Limberg, K. J. (2007). *Aircraft countermeasures (ACCM) phase II*. Kirtland Air Force Base, NM: Air Force Research Laboratory.

Task Group SAS-060. (2009). *Non-lethal weapons effectiveness assessment development and verification study* (RTO-TR-SAS-060). Neuilly sur Seine, France: North Atlantic Treaty Organisation, Research and Technology Organisation. Retrieved from https://www.cso.nato.int/pubs/rdp.asp?RDP=RTO-TR-SAS-060.

Orbons, S. (2012). Assessing non-lethal weapons use in detainee operations in Iraq: Benign force or necessary evil? *Defence Studies*, *12*(3), 452–477. doi:10.1080/14702436.2012.703847

Rahimi, R., Borve, S., & Arnesen, O. H. (2013). *Disrupting verbal communication with high-intensity directed sound.* Paper presented at the 7th European Symposium on Non-Lethal Weapons, Ettlingen, Germany.

Sauerburger, D. (1998). Make your own vision impairment simulators. *January 1998 newsletter, AER DC-Maryland chapter.* Retrieved from http://www.sauerburger.org/dona/simulators.htm

Shaughnessy, J. J., Zechmeister, E. B., & Zechmeister, J. S. (2011). *Research methods in psychology* (9th ed.). New York, NY: McGraw-Hill.

Siegel, A. B. (1991). *Eastern exit: The noncombatant evacuation operation (NEO) from Mogadishu, Somalia, in January 1991* (CRM 91-211). Alexandria, VA: Center for Naval Analyses. Retrieved from http://www.cna.org/sites/default/files/research/2791021100.pdf

Thys, N. W. (2000). *Hearing loss simulation* (Master's thesis). Cape Peninsula University of Technology, Cape Town, South Africa. Retrieved from http://digitalknowledge.cput.ac.za/xmlui/handle/11189/300

United States Marine Corps. (2011). *Mission payload module – non-lethal weapon system (MPM-NLWS) concept of operations (CONOPS).* Quantico, VA: Author.

Table 1

Metrics for assessing the task effectiveness of flashbang grenades in the noncombatant evacuation operation (NEO) scenario

Action	Measures of Effectiveness	
	In an Experiment	**From Field Data**
Climb over fence	Time until *N* participants put first foot over top of fence	Time until first demonstrator makes it to top of obstacle
Aim and throw rock	Number of “accurate” throws within time window	Time to first “accurate” throw
Aim and fire rifle	Time to first “accurate” shot	Time to first “accurate” shot
Verbally pass message	Percent of participants who accurately answer questions	Reduction in number of conversations and/or increased use of alternative communication methods (e.g., hand gesturing)

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED *(From–To)*
January 2015	Final	Oct 2014 – Jan 2015

4. TITLE AND SUBTITLE

Measures of Effectiveness for Non-Lethal Weapons:
Aligning Behavioral Experiments with Operational Success

5a. CONTRACT NUMBER
HQ0034-14-D-0001

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Shelley Cazares
James Belanich
Jeffrey A. Snyder
P. M. Picucci
Jenny R. Holzer

5d. PROJECT NUMBER

5e. TASK NUMBER
DU-2-3825

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Institute for Defense Analyses
4850 Mark Center Drive
Alexandria, VA 22311-1882

8. PERFORMING ORGANIZATION REPORT NUMBER

IDA Document NS D-5409
Log: H14-000028

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Joint Non-Lethal Weapons Directorate
3097 Range Road
Quantico, VA 22134-5100

10. SPONSOR/MONITOR'S ACRONYM(S)

JNLWD

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited (15 May 2015).

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Non-lethal weapons (NLWs) are weapons, devices, and munitions that are explicitly designed and primarily employed to incapacitate targeted personnel or materiel immediately, while minimizing fatalities, permanent injuries to personnel, and undesired damage to property in the target area or environment. To assess the technical maturity of a NLW system, combat developers compare the system's capabilities to requirements. NLW system requirements could stipulate what physiological degradation the NLW must elicit in the targeted personnel (e.g., temporary visual/hearing impairment in the case of flashbang grenades). Testing such requirements can be straightforward in the laboratory. However, physiology-based requirements can be misleading, since they do not always assess how effectively the NLW can influence the actions of the targeted personnel. Instead, NLW system requirements should, in some cases, stipulate what behavior the targeted personnel must exhibit in response to the NLW. Setting and testing behavior-based requirements is difficult, however, since many factors can influence the targeted personnel's behavior. We developed a framework to guide combat developers in setting and testing behavior-based requirements for NLWs. The framework consists of six main questions to provide structure and discipline for combat developers when determining what behavioral experiments and field data analyses are needed to assess the task effectiveness of specific NLWs in specific military missions. We exercised the framework for a Noncombatant Evacuation Operation scenario, in which U.S. forces consider deploying flashbang grenades against a crowd possibly mixed with para-military forces demonstrating hostile intent. Combat developers could repeat this analysis for other NLWs in other military missions.

15. SUBJECT TERMS

Non-Lethal Weapon, Flashbang, Noncombatant Evacuation Operation, Crowd Control, Measure of Effectiveness

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
					Dr. Shannon Foley
a. REPORT	b. ABSTRACT	c. THIS PAGE			19b. TELEPHONE NUMBER *(include area code)*
Uncl.	Uncl.	Uncl.	SAR	20	(703) 432-0916

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39.18